ПРОГНОЗИРОВАНИЕ ЗАГРУЖЕННОСТИ АВТОМОБИЛЬНЫХ ДОРОГ

Сергей Пупырев, Александр Пронченков
Уральский государственный университет, Екатеринбург
e-mail: spupyrev@gmail.com, alexander@pronchenkov.name

Аннотация

В статье описывается опыт участия авторов в конкурсе «Интернет-Математика 2010» по прогнозированию загруженности автомобильных дорог на основе данных предыдущих наблюдений. Рассматриваются различные методы прогнозирования, обсуждаются вопросы их применимости на практике. Приводятся результаты тестирования методов на реальных данных, предоставленных для конкурса.

Ключевые слова: *прогнозирование, транспортный поток, моделирование.*

1. ВВЕДЕНИЕ

Многие современные города страдают от избыточного дорожного трафика. Особенно это проявляется в будние дни, когда утром значительное количество людей стремится в офисы, а в конце дня — возвращается домой. Проблема избыточного трафика приводит к тому, что оптимальные по расстоянию маршруты, проходящие через загруженные участки дорог, часто оказываются не оптимальными по времени. А поскольку загруженность тех или иных участков является труднопредсказуемой и изменяющейся во времени величиной, то задача построения оптимального по времени маршрута оказывается трудноразрешимой.

Чтобы обнаружить проблему чрезмерных транспортных потоков, достаточно проанализировать статистику о движении транспорта в городе. Можно влиять на эти потоки, например, меняя временные интервалы светофоров. Но чтобы подобное управление было эффективным, необходимо уметь строить достоверные прогнозы.

В 2010 году, в рамках серии конкурсов «Интернет-математика», компания Яндекс предложила участникам подготовить прогноз загруженности автомобильных дорог города Москвы1. При подготовке прогноза участники могли пользоваться данными наблюдений за один месяц. Окончательный прогноз должен содержать предсказания для последнего дня месяца. В настоящей работе мы описываем наш опыт участия в этом соревновании.

Дальнейший текст организован следующим образом. В главе 2 приводится описание входных данных задачи. В главе 3 мы обсужда-

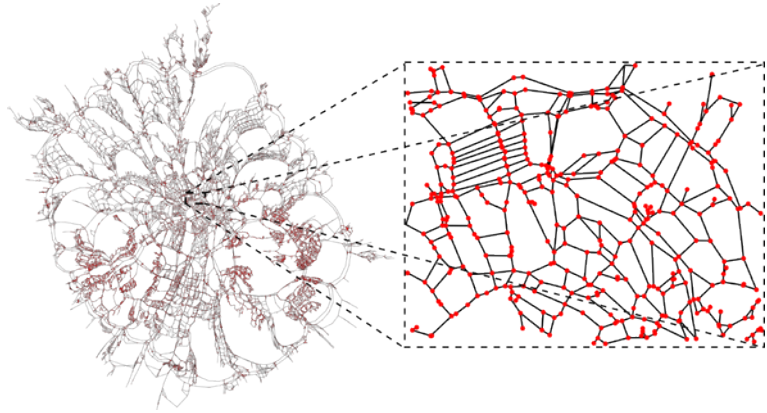

Рис. 1. Граф дорог Москвы, построенный на основе представленных данных. Для восстановления координат использовался метод многомерного шкалирования [3]

ем различные методы прогнозирования. Глава 4 содержит основные результаты, полученные нами в ходе экспериментов. В заключении мы подводим итоги наших исследований и формулируем возможные направления для дальнейшей работы.

2. ВХОДНЫЕ ДАННЫЕ

Исходные данные, предложенные для анализа в рамках конкурса, состояли из двух частей: описание сети улиц города и данные наблюдений для этих улиц. Наблюдения охватывали 31 день. Для первых 30 дней была предоставлена информация о скорости движения потока автотранспорта с 16:00 до 22:00; для последнего дня — с 16:00 до 18:00. Каждое наблюдение описывало скорость движения на отдельном участке дороги в некоторое время. Участники конкурса должны были подготовить прогноз для последнего дня с 18:00 до 22:00.

Из условий конкурса было известно, что исходные данные описывают часть города Москвы, но не были даны ни названия улиц, ни их географические координаты. Кроме того, отсутствовала информация о времени года и днях недели, когда были сделаны наблюдения. Тем не менее, часть информации можно было «восстановить» (рис. 1, рис. 3).

Предоставленные данные наблюдений оказались неоднородными: для некоторых дорог было известно много значений скоростей,

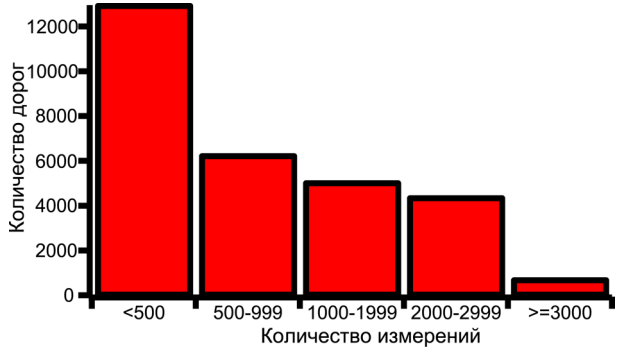

Рис. 2. Распределение данных наблюдений по дорогам

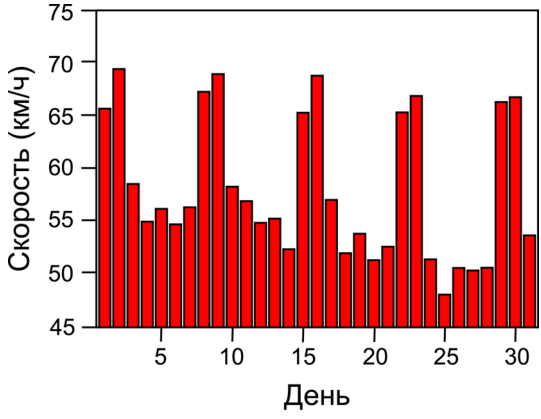

Рис. 3. Средняя скорость движения по городу по дням месяца

для других — мало. Для 55 участков нужно было построить прогноз, не имея исходных измерений по ним. На рисунке 2 изображено распределение количества данных по дорогам. Рисунок 3 содержит информацию о средней скорости движения автотранспорта в зависимости от дня месяца. На рисунке естественным образом угадывается семидневный период с характерным увеличением средней скорости в выходные дни.

На основании данных наблюдений можно сделать вывод, что средняя скорость движения по городу в каждый из дней составляет около 48 км/ч. В то же время, в исходных данных присутствовали дороги, на которых средняя скорость была меньше 10 км/ч, и дороги, где средняя скорость была больше 100 км/ч.

Для оценки прогнозов в конкурсе использовалась следующая метрика:

$$Q = \frac{1}{n}\sum k_l k_t \left| v^* - v \right| \quad (1)$$

где k_l — «коэффициент длины» (длина улицы, отнесенная к средней длине улиц (120 м)), k_t — «временной коэффициент»: 1+0.1*порядковый номер четырехминутного интервала, считая от 18:00, n — общее количество участков, v^* — наблюдаемая скорость на отдельном участке, v — прогнозируемая скорость на участке. Меньшее значение метрики соотвествует более точному прогнозу. Предсказание для длинных улиц более ценно и больший вес имеют более удаленные во времени предсказания.

Будем считать, что нам известны данные за дни $d_i \in [1..31]$ и требуется подготовить прогноз для 31-го дня. Начальный момент времени (16:00) будем обозначать через t=0, следующий момент времени, для которого известно измерение (16:02) — через t=1 и т.д. Скорость на дороге r в день d и момент времени t обозначим через $S^r_{d,t}$.

3. АЛГОРИТМЫ ПРЕДСКАЗАНИЯ

При решении задачи мы разработали и реализовали несколько методов прогнозирования. Некоторые из них работали хорошо, и они использовались для получения финального результата, другие показывали «плохие» результаты и не участвовали в формировании окончательного алгоритма. Мы не исключаем, что «плохие» алгоритмы прогнозирования можно модифицировать и подобрать к ним настройки таким образом, чтобы они давали приемлемый результат на предложенных данных. Ниже мы описываем полученный нами опыт, как положительный, так и отрицательный. Точные оценки и сравнение методов приводятся в главе 4.

3.1. Базовая модель

Как узнать, какая скорость движения будет на дороге *r сегодня* в момент времени *t*? Возникает естественное желание «подсмотреть», что в этот момент времени происходило на дороге *вчера*. У такого подхода есть две проблемы:

- в предыдущий день для данного момента времени может не быть наблюдений (например, из-за неработоспособности датчика измерения скорости), или данные такого наблюдения содержат погрешность;

• характер движения на дорогах в предыдущий день может существенно отличаться от характера движения сегодня (например, из-за дорожных работ или погодных условий вчера на некотором участке была пробка, а сегодня пробки нет).

Первую проблему можно частично преодолеть, если вместо вчерашнего наблюдения в строго тот же момент времени t, рассматривать усреднённые данные за небольшой промежуток $[t-\delta, t+\delta]$. Вторая проблема становится менее заметной, если учитывать в предсказаниях не только вчерашний день, а все предыдущие дни. Таким образом, мы приходим к следующему выражению:

$$S^r_{d,t} = Average\left(S^r_{d-1,t}, S^r_{d-1,t-1}, ..., S^r_{d-1,t-\delta}, S^r_{d-1,t+1}, ..., S^r_{d-1,t+\delta}, S^r_{d-2,t}, S^r_{d-2,t-1}, ...\right) \quad (2)$$

где среднее значение вычисляется по всем доступным измерениям из интервала $[t-\delta, t+\delta]$ для всех предыдущих дней $[1, d\text{-}1]$. Прогноз, построенный по формуле (2), назовем *базовой моделью I*. Наилучших результатов мы добились при использовании значений $\delta \in [30,60]$, т.е. каждое значение скорости вычисляется на основе, как минимум, часового интервала.

Несмотря на свою простоту, модель I строит достоверный прогноз и, при достаточной полноте исходных данных (показаний скоростей без «пропусков» за длительный промежуток времени), может быть успешно использована без модификаций. Однако предоставленные в рамках конкурса данные имели существенные пробелы.

Если для некоторой дороги имеется малое количество измерений, то значение, вычисленное по формуле (2), будет существенно зависеть от шума в исходных данных, и прогноз будет неточным. Поэтому мы использовали следующую модификацию:

$$S^r_{d,t} = \frac{AvgValue \cdot AvgCount + GlobalAverage \cdot K}{AvgCount + K}$$

где *AvgValue* — значение скорости, полученное на основе формулы (2); *AvgCount* — количество наблюдений, используемых в модели I; *GlobalAverage* — средняя скорость в момент времени t по всем дорогам и всем предыдущим дням, $K>0$ — параметр модели. В такой модификации, если имеется достаточное количество исходных данных (т.е. *AvgCount* велико), то результат прогноза будет близок к значению *AvgValue*. Если же данных недостаточно, то прогноз корректируется в сторону *GlobalAverage*, которое в этом случае является более вероятным ответом. В нашей реализации использовались значения параметра $K \in [15,100]$.

Что понимается под функцией *Average* в выражении (2)? Арифметическое среднее, геометрическое, какое-то другое? Ответ на этот вопрос дает вид формулы (1) для оценки качества результата, в которой используется L_1 метрика. Предположим, для предсказания у нас есть измерения скорости $s_1,s_2,...,s_k$, которые мы будем подставлять в формулу (2). Без ограничения общности считаем, что $s_1 \le s_2 \le ... \le s_k$. Легко понять, что оптимальным значением для предсказания с точки зрения оценки (1) является *медианное* среднее значений $s_1,s_2,...,s_k$, т.е. $Average(S_1,S_2,...,S_k):=(S_{\lfloor (k+1)/2 \rfloor}+S_{\lfloor (k+1)/2 \rfloor})$. Алгоритм с использованием медианного среднего назовем *базовой моделью II*.

Следующий шаг оптимизации основан на использовании различий в состояниях дорожного трафика в разные дни недели. В модели II все дни недели учитываются с одинаковым весом, и, например, измерения, сделанные в понедельник, влияют на результат так же, как «пятничные» измерения. Кроме того, не учитываются, погодные условия: если в один из предыдущих дней выпал снег, то скорости в этот день в среднем по городу будут ниже, и это может негативно сказаться на предсказаниях в ясный бесснежный день. Мы использовали следующую схему. На первом шаге вводится *метрика похожести* $\omega(d_i,d)$ каждого дня $d_i \in [1..d-1]$ на последний день месяца d. Данная метрика показывает насколько транспортные потоки в день d_i похожи на потоки в день d. Для последнего дня месяца известны значения скоростей только для периода времени [16:00; 18:00), поэтому метрика вычисляется на основе этого периода. На втором шаге модифицируется формула (2) так, чтобы она учитывала метрику похожести. При этом измерения похожих дней оказывают более сильное влияние на результат, чем измерения, сделанные в непохожие дни. Метрику похожести мы вычисляли по следующей формуле:

$$\omega(d_i, d) = \left(\frac{1}{|R|} \sum_{r \in R} \sum_{t \in T} \frac{|S^r_{d,t} - S^r_{d_i,t}|}{|T|} \right)^{-1}$$

где R — множество всех дорог, T — множество моментов времени, для которых известны оба измерения $S^r_{d,t}$ и $S^r_{d_i,t}$ (это множество зависит от дороги r). Таким образом, если все измерения скоростей в день d_i совпадают с измерениями скоростей в те же моменты времени в день d, то $\omega(d_i,d)=\infty$. Напротив, если дни d_i и d существенно различаются, то $\omega(d_i,d)$ принимает значения, близкие к нулю. На практике мы получили $0.03 \le \omega(d_i,d) \le 0.1$.

Далее выражение (2) модифицируется так, чтобы в ней учитывались найденные значения похожести. Опять предположим, что у нас есть набор измерений для одной дороги $s_1,s_2,...,s_k$ (без ограничения общности, $s_1 \le s_2 \le ... \le s_k$), по которым строится предсказание. Этой последо-

вательности сопоставляется последовательность весов $\omega_1,\omega_2,...,\omega_k$, где ω_i — это похожесть между днем d и днем, в котором сделано измерение s_i. Веса ω_i нормализуются так, чтобы выполнялось $\sum_{i=1}^{k} \omega_i = 1$ (для этого достаточно разделить каждый элемент последовательности на сумму всех элементов). Оптимальным значением предсказания в метрике L_1 является взвешенная медиана, т.е. s_{opt}, для которого выполняется $\sum_{i=1}^{opt} \omega_i < 0.5$ и $\sum_{i=1}^{opt+1} \omega_i \geq 0.5$. Модель с использованием взвешенной медианы будем называть *базовой моделью III*.

3.2. Сингулярное разложение

Сингулярное разложение матриц (singular value decomposition, далее SVD) — часто используемый метод в задачах, когда необходимо «предсказать» неизвестные элементы матрицы. Наряду со смежным методом главных компонент (principal component), сингулярное разложение часто и успешно используется для предсказаний дорожного трафика.

Первый шаг алгоритма — построение матрицы M скоростей для каждой дороги; на пересечении строки i и столбца j записывается значение скорости в момент времени i, измеренное в день j. Исходные данные неполны, поэтому построенная матрица будет иметь пропуски. Кроме того, в матрице M неизвестны элементы, соответствующие моментам времени, для которых нужно сделать предсказания (рис. 4).

Для построенной таким образом матрицы M размера $n \cdot m$ подбираются две матрицы U и V размером $n \cdot f$ и $m \cdot f$ соответственно (при этом используется сингулярное разложение матрицы M, откуда и происходит название метода), где число f называется размерностью

Рис. 4. Матрица M скоростей в методе SVD

метода SVD. Матрицы U и V строят так, чтобы выполнялось

$$m_{ij} \approx \sum_f u_{if} \cdot v_{jf}$$

для всех известных элементов m_{ij}. После этого по той же формуле можно оценить («предсказать») пропущенные элементы m_{ij}.

Мы реализовали несколько вариаций метода SVD, отличающихся методом нахождения матриц U и V, размерностями (f=10,20,40,80) и способом формирования исходной матрицы M (в частности, M можно «расширить», записывая в строки информацию о группе дорог). При этом ни одна из модификаций не дала приемлемого результата. Наилучшие полученные прогнозы сравнимы с базовой моделью II и значительно проигрывают базовой модели III. Мы предполагаем, что этому есть две причины: (а) исходные данные содержат большое количество шума, мешающего найти «правильное» разложение, и (б) исходных данных оказалось недостаточно для метода SVD.

3.3. Нейронные сети

Нейронные сети, казалось бы, являются идеальным инструментом для решения задач прогнозирования. Мы потратили существенное количество времени на выбор подходящей топологии сети, способа ее обучения, вида входных и выходных данных. Нам так и не удалось научиться получать конкурентоспособные результаты при помощи нейронных сетей. Наиболее адекватные оценки получались при следующих параметрах:

- 4-6 скрытых слоев перцептронов;
- входные данные представляют собой вектор из скоростей в моменты времени t,t+1,t+2,...,t+k-1, где k из диапазона [6, 10];
- на выходе величина скорости в момент времени t+k;
- обучение с использованием метода обратного распространения ошибки;
- обучение сети происходит для каждой дороги независимо.

3.4. Предсказания на основе соседей

Как упоминалось выше, исходные данные содержат большое количество «пропусков», что затрудняет использование других методов. Наша идея состоит в том, чтобы восстановить недостающие измерения, используя информацию о трафике на соседних дорогах. Зафиксируем некоторую дорогу r и ее соседей N (ребра в графе дорог на расстоянии, не превышающем фиксированного значения). Гипотеза состоит в том, что скорость на дороге r в каждый момент

времени является некоторой функцией от скоростей на N в тот же момент времени:

$$S^r_{d,t} = f\left(S^{r_1}_{d,t}, S^{r_2}_{d,t}, ..., S^{r_N}_{d,t}\right)$$

В качестве f мы выбирали линейную функцию (т.е. использовали линейную регрессию), вид которой восстанавливался при помощи метода наименьших квадратов. Обучение проходило на тех t, для которых известны все скорости $S^r_{d,t}, S^{r_1}_{d,t}, S^{r_2}_{d,t}, ..., S^{r_N}_{d,t}$, а восстановление координат — для тех t, для которых не было известно $S^r_{d,t}$, но были известны скорости всех соседей $S^{r_1}_{d,t}, S^{r_2}_{d,t}, ..., S^{r_N}_{d,t}$. Мы также отбрасывали все «ненадежные» предсказания — те, для которых функция f строилась на основе менее чем 30 точек.

Данный подход оказался полезным на практике. В частности, количество исходных данных возросло примерно в 1.5 раза. Прогнозы, построенные с помощью базовой модели на основе новых данных, оказались более точными.

3.5. Общие модификации и усовершенствования

3.5.1. Кластеризация

Характер движения автотранспорта по различным дорогам неоднороден. На магистралях движение в течение всего дня имеет приблизительно постоянную скорость. То же самое справедливо для улиц в спокойных спальных районах. Напротив, в центре города загруженность выше, пробки возникают в конце рабочего дня и постепенно ослабевают к вечеру. Нам кажется естественным применять различные методы предсказаний для разных групп дорог и разного времени суток. Мы разработали несколько схем кластеризации дорог. Для каждого построенного кластера использовались разные модели или разные параметры моделей. Для получения финального результата мы использовали следующую схему кластеризации:

- дороги, для которых было дано малое количество измерений (в нашей реализации — менее 500 за весь период), обрабатывались согласно базовой модели II;
- все остальные дороги были разделены на три группы: (1) дороги с «постоянной скоростью», для которых стандартное отклонение значений скорости лежит в диапазоне от 0 до 10; (2) дороги со стандартным отклонением в диапазоне от 10 до 20; (3) дороги с «большим количеством пробок», с отклонением более 20. Для каждой из таких групп использовалась независимая базовая модель III;

• использовалась кластеризация измерений по времени: все модели независимо обучались для промежутков времени [18:00-20:00) и [20:00-22:00).

3.5.2. Глобальные эффекты

Предварительный этап обработки, предшествующий любому нашему методу, заключался в удалении данных, соответствующих выходным дням. Мы обнаружили, что используемые нами методы дают лучшие результаты на «очищенных данных». Это вполне естественно, поскольку характер движения в выходные дни существенно отличается от движения в будни.

Внимательное изучение рисунка 3 приводит к следующей мысли: средняя скорость по городу с течением времени уменьшается. Это может быть сезонный эффект или возрастающее количество автотранспорта на дорогах. В любом случае полезно «удалить» этот глобальный эффект перед построением прогнозов. Перед началом обработки данных все скорости $S^r_{d,t}$ были изменены на значение $S^r_{d,t}+k{\cdot}d$, где k — параметр, отвечающий за изменение средней скорости по всем дорогам в течение месяца. В зависимости от группы дорог $k\in[0.01..0.02]$. Легко убедиться, что в новых «повернутых» измерениях средняя скорость движения остается постоянной. Перед выводом ответа необходимо поменять результат на $S^r_{d,t}-k{\cdot}d$.

Другой полезный трюк, который мы использовали на этапе предварительной обработки данных, — это нормализация скоростей, при которой из каждого измерения $S^r_{d,\,t}$ вычитается среднее значение скорости на этой дороге за весь промежуток времени. При этом некоторые из $S^r_{d,\,t}$ могут стать отрицательными, а их сумма по всем парам d,t будет равна нулю. Интуитивное обоснование этой операции следующее: пара дорог может иметь сильно отличающиеся абсолютные значения (например, из-за разных ограничений скорости), но при этом иметь похожие колебания скорости относительно своего среднего значения. Особенно полезной эта операция оказалась для метода SVD.

3.5.3. Комбинирование различных подходов

Используя опыт участников соревнования по прогнозированию рейтингов фильмов Netflix (http://www.netflixprize.com), в котором наилучшие результаты принесло комбинирование принципиально различных алгоритмов, мы пробовали объединять результаты используемых нами моделей. Пусть $X_1,X_2,...,X_k$ — это вектора,

элементами которых являются предсказания скоростей каждой из k моделей. Размер вектора X_i совпадает с количеством искомых предсказаний. Мы хотим построить линейную комбинацию $X_1\alpha_1+X_2\alpha_2+...+X_k\alpha_k$ так, чтобы результирующий вектор был наиболее близок к ответу. Чтобы найти коэффициенты линейной регрессии $\alpha_1,\alpha_2,...,\alpha_k$, мы делаем предсказания $Y_1,Y_2,...,Y_k$ для одного из предыдущих дней, для которого нам известен правильный результат Y. После этого решается задача Least Absolute Deviations [5]: $\min||Y_1\alpha_1+Y_2\alpha_2+...+Y_k\alpha_k-Y||_1$. Полученные коэффициенты регрессии $\alpha_1,\alpha_2,...,\alpha_k$ можно использовать для комбинирования $X_1,X_2,...,X_k$.

В наших экспериментах мы объединяли около 30 различных результатов, среди которых были построенные с использованием базовых моделей, нейронных сетей, SVD и их вариаций с разными параметрами. Коэффициенты регрессии мы искали на 24-м и 26-м днях, поскольку именно они наиболее похожи на предсказываемый 31-й день. Комбинирование всех моделей без базовой модели III дает существенный выигрыш и заметно улучшает итоговую оценку. Однако, присоединение базовой модели III полностью меняет картину: коэффициент для этой модели оказывается равным ≈0.95 (а все остальные в сумме дают оставшиеся 0.05), и итоговый результат практически совпадает с базовой моделью.

Наша модель на основе взвешенной медианы оказалась намного лучше остальных и ее вклад «перевесил» вклад остальных моделей. Тем не менее, мы полагаем, что комбинирование нескольких методов может улучшать прогноз. Для этого нужно иметь несколько равноценных алгоритмов.

4. РЕЗУЛЬТАТЫ

Результаты тестирования основных алгоритмов представлены на рисунке 5 (на момент написания статьи мы не имели доступ к данным по 31-му дню. Поэтому все приведенные оценки рассчитаны для 26-го дня. Для других дней конкретные значения отличаются, но их отношение сохраняется). Отметим, что предложенная компанией Яндекс метрика (1) для оценки качества предсказания неинтуитивна: глядя на результат, сложно сказать, насколько хорошо или плохо работает алгоритм. Например, неясно, что означает оценка 58. Мы приводим в качестве альтернативной оценки среднее отклонение значений прогноза от наблюдаемых измерений (рис. 6). Нельзя сказать, что такая оценка лучше исходной, но она отражает более

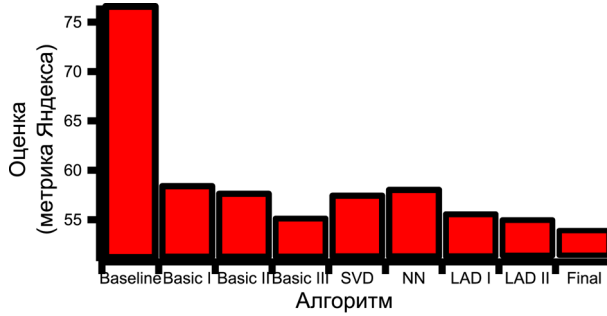

Рис. 5. Результаты тестирования алгоритмов по метрике Яндекса

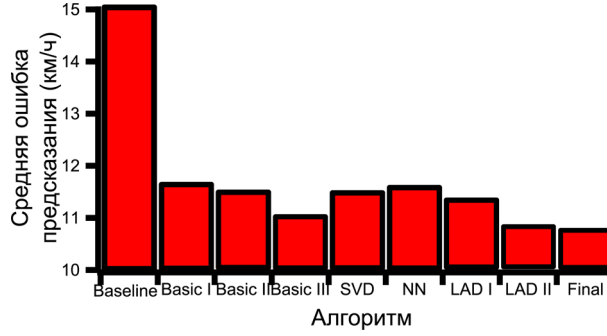

Рис. 6. Среднее отклонение предсказания скорости

интуитивное представление о точности метода. Используемые на рисунках обозначения алгоритмов следующие:

Baseline — это «простая скептическая оценка»: средняя скорость для участка по всем дням месяца в одно и тоже время; если данных не было, то мы считали скорость равной 0;

Basic I,II,III — базовые модели I,II и III соответственно;

SVD — алгоритм, на основе метода сингулярного разложения;

NN — алгоритм, на основе нейронных сетей;

LAD I — комбинирование результатов вышеупомянутых алгоритмов, за исключением базовой модели III;

LAD II — комбинирование результатов базовой модели III и LAD I;

Final — наш финальный результат, построенный на основе базовой модели III с использованием кластеризации.

Проанализировав результаты, мы сделали следующие выводы:

1. Сложные модели (SVD, нейронные сети) требуют больших трудозатрат при реализации и более тщательного подбора параметров, чем простые модели, основанные на подсчете средних значений за предыдущие дни наблюдений. При этом на практике они дают сравнимые результаты.

2. Наилучшие результаты нам принесли две идеи: (1) использование различий между днями и переход ко взвешенной медиане (сравните Basic II с Basic III на рис. 5); (2) кластеризация и учет глобальных эффектов (переход от Basic III к Final). Отметим также «силу» метода комбинирования различных подходов. Объединение нескольких результатов (Basic I, Basic II, SVD, NN и их модификаций) дает почти тот же результат, что Basic III модель.

3. Самая простая модель (Basic I) в среднем ошибается на 11.66 км/ч, тогда как лучшая даёт среднюю ошибку в 10.78 км/ч, т.е. улучшение составляет менее 1 км/ч. Мы оставляем открытым вопрос об актуальности разработки «умных» алгоритмов для предсказания загруженности трафика.

4. Самый сложный период времени для прогнозирования — с 19 до 20 часов, для него средняя ошибка составляет 11.28 км/ч. Самый простой период времени — с 20 до 21 для которого средняя ошибка 9.8 км/ч. Для сравнения: ошибка в период с 18 до 19 составляет 10.99 км/ч, с 20 до 21 — 10.8 км/ч.

5. ОБЗОР СУЩЕСТВУЮЩИХ ПОДХОДОВ

Список литературы по теме прогнозирования загруженности транспортных сетей обширен. До недавнего времени существовал специализированный журнал The Journal of Transportation and Statistics (JTS), в котором публиковались последние достижения из этой области. Ежегодно проходит научная конференция World Congress on ITS1, на которой имеется секция, посвящённая исследованиям дорожных сетей.

Отметим несколько работ, на которые мы опирались при создании своих моделей. Работы [1] и [2] описывают методы понижения размерности данных на основе разложения матриц. В статье [4] рассмотрен вопрос о построении линейных регрессионных моделей по предыстории наблюдений. В этой работе отмечается, что такие модели обладают хорошими прогностическими свойствами. Нейронные сети и аспекты их применения для прогнозирования загруженности дорог описаны в работе [6]. Модели на основе равновесия потоков в сети можно найти в работе [9].

В работе [7] приводится анализ задачи прогнозирования дорожного трафика. Авторы этой работы разделяют прогнозы на краткосрочные и долгосрочные. При построении краткосрочных прогнозов важным является анализ мгновенных состояний транспортного потока, а в долгосрочных прогнозах больший вес приобретают социальные и экономические тенденции. Кроме того, различаются два типа прогнозов: прогноз скорости движения транспортного потока и прогноз величины потока, т.е. количества автомобилей которые преодолевают участок дороги в единицу времени.

Сложность создания универсальных методов прогнозирования для различных внешних условий обсуждается в работе [8]. В этой работе говорится о том, что на практике часто используются несколько независимых методов, каждый из которых ориентирован на определённые внешние условия, такие как погода, время суток, день недели, время года.

Отметим общий недостаток изученных нами работ. В этих работах почти полностью отсутствует сравнение новых результатов с опубликованными ранее; и в каждой из работ используется собственный набор данных для анализа. В результате оказывается невозможным сравнение методов разных авторов на основе только текстов публикаций. Вероятно, причиной тому является отсутствие общедоступного набора данных наблюдений за продолжительный период времени.

6. ЗАКЛЮЧЕНИЕ

В настоящей работе мы рассмотрели несколько методов для построения прогнозов загруженности дорог на основе предыдущих измерений. Благодаря конкурсу, организованному компанией Яндекс, мы протестировали предложенные алгоритмы на реальных данных. Основной вывод наших исследований — простые алгоритмы, основанные на вычислении средних скоростей, работают лучше алгоритмов, построенных на основе сложных моделей. В частности, наш алгоритм не использует граф дорог и обрабатывает каждую дорогу независимо от остальных.

Среди возможных направлений дальнейших исследований по теме отметим следующие задачи. Применимы ли предложенные алгоритмы для других исходных данных? Достаточно ли было исходных данных для работы метода и построения соответствующих оценок? Существуют ли иные эвристические алгоритмы, позволяющие получать значительно более точные прогнозы? Также интересен вопрос, нужны ли дополнительные данные о дорогах (граф дорог, длина улиц,

ограничения скорости и т.д.) или достаточно истории измерений скорости за большой промежуток времени. Являются ли прогнозы, учитывающие такую информацию, более точными?

7. БЛАГОДАРНОСТИ

Авторы выражают благодарность В. Канторову и А. Коковину за содействие в работе, плодотворные обсуждения и реализацию описанных алгоритмов.

ЛИТЕРАТУРА

1. **Tsekeris, T. Stathopoulos, A.** Measuring variability in urban traffic flow by use of principal component analysis. Journal of Transportation and Statistics. 9 (1), pp. 49–62, 2006.
2. **Nishiuma, N., Goto, Y., Kumazawa, H., Nikovski, D., Brand, M.**, Traffic Prediction using Singular Value Decomposition. World Congress on ITS, 2004.
3. **Gansner, E., Koren, Y., North, S.**, Graph Drawing by Stress Majorization. Proceedings of 12th Int. Symp. Graph Drawing, Lecture Notes in Computer Science, 3383, Springer-Verlag, pp. 239–250, 2004.
4. **Sun, H., Liu, H., Xiao, H., He, R., and Ran, B**. Short Term Traffic Forecasting Using the Local Linear Regression Model. Journal of Transportation Research Board, 1836, pp. 143–150, 2003.
5. **Schlossmacher E.J.**, An iterative technique for absolute deviations curve fitting. Journal of the American Statistical Association 68, pp. 857-859, 1973.
6. **Yasdi, R.**, Prediction of road traffic using a neural network approach. Neural Computation and Application, v8, pp. 135-142, 1999.
7. **Smith, B., Williams, B. and Keith, O.**, Comparison of parametric and nonparametric models for traffic flow forecasting, Trans. Res. C, vol. 10, no. 4, pp. 257–321, Aug. 2002.
8. **Min, W., Wynter, L., Amemiya, Y.**: Road traffic prediction with spatio-temporal correlations. In: Proceedings of the Sixth Triennial Symposium on Transportation Analysis, Phuket Island, Thailand, 2007.
9. **David, B.**: A Practitioner's Guide to Urban Travel Forecasting Models, 12th International EMME/2 Users Conference, 1997.